COVID-19 Outbreak Detection Tool

Methodology

Last updated: June 24, 2021

COVID19sim@mgh.harvard.edu

Massachusetts General Hospital Institute for Technology Assessment
Harvard Medical School
101 Merrimac St. STE 1010 • Boston, MA 02114 • United States

Model Overview

The *COVID-19 Outbreak Detection Tool* is designed to detect recent county-level COVID-19 outbreaks and predict how fast an outbreak would spread in each county in the United States. Using a machine learning approach, the tool estimates the doubling time of COVID-19 cases in each county by accounting for reported COVID-19 cases, COVID-19 deaths, face mask mandate, social distancing policies, social vulnerability index of each county (e.g. income level, employment rate), as well as the daily COVID-19 test and vaccination ratios of each county. The tool is updated at least once a week.

Methodology

We built a machine learning based generalized random forest model to estimate the growth rate of incident COVID-19 cases in each county, as defined below:

$$x(t_0 + t) = x(t_0)e^{rt},$$

where $x(t_0)$ is the current incident case number, and $x(t_0 + t)$ is the incident case number *t* days later if the exponential growth rate *r* persists.[1] In our heat map, we convert this exponential growth rate to a more intuitive concept of case doubling time, i.e., if an exponential growth rate *r* persists, the initial case number would double in $\frac{\ln(2)}{r}$ days.

The incident case number at a time point $x(t)$ is defined as:

$$x(t) \ = I(t) - I(t - 22)\,,$$

where $I(t)$ and $I(t - 22)$ are the cumulative infectious case number up to time t and $t - 22$, respectively. We assumed that a patient is either recovered or deceased 22 days after getting infected.[2] We used the 7-day moving average to smooth out the weekend effect of case number oscillations.[3] Furthermore, our analysis excludes counties with less than 20 incident cases at time t to reduce noise from the data.

To obtain robust and up-to-date county-level exponential growth rate estimations, we followed the below steps.

1. We first constructed a database of daily COVID-19 case number growth rates for each county by utilizing the historical cumulative confirmed case numbers reported by the New York Times (https://github.com/nytimes/COVID-19-data).

2. We then augmented these growth rates with various features capturing relevant factors affecting the disease spread. This includes county specific characteristics such as the geographic, demographic and social vulnerability index, along with time specific characteristics, such as economic stimulus packages and social distancing interventions. The following features were considered in our analysis:
    - Historical growth pattern and cumulative confirmed case number in each county:
        - daily case growth rate throughout the history
        - initial cumulative case number for each day
    - Geographical location of each county, as captured by the county centroid longitude and latitude
    - Social vulnerability index (SVI) of each county such as (see the SVI data dictionary for a complete list):

    - per capita income
    - employment rate
    - insurance coverage
- COVID-19 related economic and social distancing policy of each state such as (See the CUSP data dictionary for a complete list):
    - face mask mandate
    - mandate quarantine for those entering the state
    - paid sick leave
- COVID-19 testing:
    - Test positivity ratio
- COVID-19 vaccination rate:
    - Proportion of population that has received partial dose of vaccine
    - Proportion of population that has received full dose of vaccine

3. Finally, we used the generalized random forest algorithm[4] to match the incident case number growth trends based on these features and estimated an exponential growth rate for each growth pattern. This method allows us to balance the *bias-variance tradeoff* in exponential growth rate estimations.
    a. First, our estimations are based on the most recent growth pattern for each county, which accounts for changing conditions and policies in a county and thus reduces *bias*. For example, the case number growth patterns before a mask mandate should not be used to estimate exponential growth rate after the mandate because it will confound the model.
    b. Second, this method pools together all relevant trends across counties and throughout the history in our exponential growth rate estimation for each county, which reduces the *variance* of our estimates. For example, to estimate the exponential growth rate of county A, the algorithm is able to use data from a different county B to reduce the estimation variance, provided that county A and B are sufficiently similar in various county- and state-level features.

Implementation and analysis were performed in R.

References

1. Ma J. Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling.* 2020;5:129-141.
2. The University of Melbourne. Coronavirus 10-day forecast. *Available at* *https://COVID19forecastscienceunimelbeduau/.* August, 26, 2020.
3. Bergman A, Sella Y, Agre P, Casadevall A. Oscillations in US COVID-19 incidence and mortality data reflect diagnostic and reporting factors. *Msystems.* 2020;5(4).
4. Athey S, Tibshirani J, Wager S. Generalized random forests. *The Annals of Statistics.* 2019;47(2):1148-1178.

Data Sources

- Historical Growth Pattern and Cumulative Confirmed Case Number of each county:
  Source: The New York Times (https://github.com/nytimes/COVID-19-data)

- Geographical Location of each county:
  Source: 2019 U.S. Census Gazetteer Files (https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html)

- Social Vulnerability Index (SVI) of each county:
  Source: Centers for Disease Control and Prevention Social Vulnerability Index 2018 Database (https://svi.cdc.gov/data-and-tools-download.html)

- COVID-19 related Economic and Social Distancing Policy of each state:
  Source: COVID-19 US state policy database (CUSP)
  (https://docs.google.com/spreadsheets/d/1zu9qEWI8PsOI_i8nI_S29HDGHlIp2lfVMsGxpQ5tvAQ/edit#gid=1643322116)

- COVID-19 testing rate and vaccination rate of each county:
  Source: COVID Act Now
  (https://apidocs.covidactnow.org/data-definitions)